1902 YOUNG — Adjustment of Census Age returns.

THE ADJUSTMENT OF CENSUS AGE RETURNS[1].

In order to put the problem before us in its simplest form it will be well to conceive the ages returned at an enumeration of the population as arranged in a table of two columns. The first of these columns, which may have the heading "x", contains in order the possible ages of human life, expressed in years, from zero to the maximum age. Opposite each of these ages and in the second column is placed the number of persons who have reported themselves as of that age. The second column may be headed "P_x".[2] Thus P_{24} represents the number reported as 24 years old. It is evident that in this table the value of x progress from 0 to 100+ in simple arithmetical progression. But the value of P_x decreases as x increases, approaching zero at the maximum value of x. This decrease, however, is not constant. P_x is not always smaller than P_{x-1} or P_{x-2}.

Consideration of the forces which influence the age constitution of the population shows that there is a necessary relation between the age and the number living at that age. This may be shown by assuming a hypothetical population in which there is no relation between the values of x and P'_x (when P'_x represents the true number living at age x). In such a population P'_x might be constant.

1 This is the third of a series of three studies in age statistics. The first, on "The Comparative Accuracy of Different Forms of Quinquennial Age Groups." appeared in the Quarterly Publications of the American Statistical Association for March, 1900, and was devoted to an analysis of the nature of the errors in the reported ages of adults. The second article, on "The Enumeration of Children," was printed in the same Publications for March, 1901, and dealt with the errors in the reported ages of children, with special reference to the relation of those errors to the apparent deficiencies in the number of children enumerated in the census. In connection with each of these studies I am under special obligations to Professor Walter F. Willcox, of Cornell University, Chief Statistician for the Division of Methods and Results of the Twelfth United States Census.

2 Borrowing a term sometimes used in life table notation to designate the mean number of survivors in any year of life.

Now the force which in an actual population is most effective in causing variations from this hypothetical condition—the force of mortality—has been found to be so regular in its action that it may be very closely represented by a mathematical formula in which it is made a function of the age.[1]

The other causes of divergence from the hypothetical constant value of P'_x are: (a) variations in the force of mortality in different years, (b) variations in the birth-rate, (c) migration. Our knowledge of the forces behind these disturbing factors is incomplete, but may be in part supplied by empirical knowledge about the factors themselves. In short, the number living may be properly considered a function of the age, the form of the function being determined by the factors just considered. While the exact form of the function must be unknown, an examination of the nature of the determining factors will often enable us to discover whether a reported number P_x can correctly represent the value of P'_x.

If the values of P_x, as returned by any census, be studied, it will be seen that the irregularities in the series seem to follow a more or less definite law. This may be seen most clearly in a graphic representation of the population, classified by the reported ages. If on the axis of abscissas we erect successive ordinates, in such a manner that the area included between the ordinates at x and x+1 represents P_x it will be seen that for ages over 20, the values of P_x, in addition to following the general order of decreasing as the values of x increase, are relatively larger at each recurrence of certain forms of the value of x. The values of x may be classified in the order of the corresponding values of P_x as follows:

(1) Numbers whose last digit is 0.
(2) Numbers whose last digit is 5.
(3) Numbers whose last digit is 2, 4, 6 or 8.
(4) Numbers whose last digit is 1, 3, 7 or 9.

[1] Gompertz' Law of Mortality as modified by Makeham. Cf. below, p. 98.

In those censuses in which the age question relates to date of birth, rather than to the number of years lived, the same order holds good, if we consider x as representing the year of birth.

None of the different factors which affect the values of P'_x are of such a nature as to produce these regularly recurring relative maxima and relative minima. We must conclude that the P_x series does not truly represent the P'_x series. Hence the need of adjustment.

The legitimacy of applying some method of adjustment to the census age tables hinges on several considerations, chief among which is the greater probability of the P'_x series forming a fairly smooth progression than of its conforming to the irregularities of the P_x series. This probability is supported by both analysis and experience. As has been stated above, none of the factors which determine the form of the function P'_x is of such a nature as to produce sudden irregularities in the series. It is certain that changes in the birth and death rates as well as variations in the amount and direction of migration must leave their marks in the form of certain irregularities. On this account it is probable that the curve representing the P'_x series contains several points of inflexion [1] and that the first differences of the series do not always have the same sign.[2]

Yet these flexures must not be confounded with the sharp angles of the curve representing the P'_x series. Without going into further details, we may conclude that it is not probable that in the age constitution of the population the principle of continuity is so far violated that the P_x series can not be accurately represented by a fairly smooth curve. This conclusion, based on analysis of the controlling factors, is confirmed by experience. The greater the accuracy of a census, arising from special care in the enum-

1 A point of inflexion occurs when a curve changes from concave towards the abscissa to convex, or *vice versa*.

2 That is, that while the values of P'_x generally decrease as x increases, in certain parts of the series the values of P'_x increase with the values of x.

eration or from the intelligence of the persons enumerated, the more nearly does the curve representing the age returns approach smoothness of form. Hence the legitimacy of applying a curve smoothing process to the P_x series.

Before passing on the subject of methods, it will be well to consider briefly the tests of a good adjustment. The thing desired is a smooth series which will adhere as closely as possible to the facts. The aim should be to eliminate the irregularities in the P_x series caused by mis-statement of age, while retaining those corresponding to actual irregularities in the age constitution of the population. In testing a particular method of adjustment we shall be aided by the fact that the real irregularities usually take the form of flexures in the curve covering a period of several terms of the series, while the irregularities caused by errors are more likely to appear as angular deflections, corresponding to abnormal values of single terms. It follows that groups of terms are likely to be more accurate than single terms. Especially is this the case when our knowledge of the form of error is sufficient to enable us to choose groups so constituted that the probability of the equality of positive and negative errors is a maximum. To obtain the closest agreement with the facts such groups should contain as few terms as considerations of accuracy permit. The agreement of corresponding groups of the terms of the adjusted and unadjusted series may be considered as one test of a good adjustment.

The various methods which have been used for the adjustment of age returns and for similar purposes can be classified for purposes of examination as: (1) Methods of Substitution, (2) Arbitrary Methods, (3) Methods of Averaging, (4) Methods of Algebraic Interpolation, (5) Graphic Methods.

I. Methods of Substitution.

Speaking accurately, what we have called "methods of substitution" are not methods of adjustment. Their distinguishing feature is the discarding of the census figures and the substitution of presumably more accurate figures gleaned from other sources. I have been able to find only one instance in which a complete table of the ages of a population has been prepared in this way. The age returns of the Italian census of 1881 were subjected to careful analysis, the results of which were published as a volume of the 'Annali di Statistica.'[1] Several adjusted age tables were prepared, in one of which the census returns were not used, except for the sum total of the population.[2] The arithmetic mean of the number of registered births in 1880, 1881 and 1882 was used as the basis of the calculation. Then, on the somewhat naïve assumption that the annual number of births had been increasing in arithmetical progression, the total number of births in the decade 1863-1872 was subtracted from the number in the decade 1873-1882, the difference was divided by 10, and the quotient was called the mean annual increase. The number of births in each of the hundred years preceding the census was thus easily computed, and the probable number of survivors at the date of the census was obtained by the use of Rameri's Italian life table.[3] The total number of males thus estimated to be living at the date of the census (13,822,447) was 442,936 less than were enumerated in the census. The deficiency was ascribed to excess of emigration and other disturbing factors. The estimated numbers were therefore increased proportionately.

1 'Sulla Composizione della Popolazione per Età.' Rome, 1885. These studies were made in the 'Ufficio matematico della Statistica' under the direction of Luigi Perozzo.

2 Op. cit., p. 5. and ff.

3 The mean of the male births in 1880, 1881 and 1882 was 532,111. The mean annual increase in the number of births was found to be 2,215. Then if S_x be the number of survivors at a given age out of 1,000,000 births, by the life table, the number of males X years old living at the date of the census would be

$$532{,}111 \frac{S_x}{1{,}000{,}000} \left(1 - \frac{2{,}215}{532{,}111} X\right).$$

It need not be pointed out that the method just described is extremely crude. For the Italian population, which has experienced a fairly regular rate of increase, it seems to have given passably good results. For a population subject to greater fluctuations in the rate of growth this method would give results very far from the truth. Even in the Italian case the goddess of chance must have been propitious.

There is less difficulty in substituting figures based on the registration of births and deaths for the earlier part of the age table. This was done for ages under 10 in the Italian studies mentioned above, and the results were incorporated in tables in which the higher ages were adjusted by other methods.[1] In the census of England and Wales adjusted ages under five are obtained by distributing the number reported by the census among the single years in accordance with the registered births and deaths.[2]

In a number of censuses in which no adjustment of the returned ages is attempted, the registration reports are made to furnish the material for tables showing the probable number of children of different ages living at the date of the census.[3] Comparisons of such tables and the census age tables often lead to the detection of errors in both.

The process of substitution at the lower ages is especially useful because none of the methods of adjustment which seem especially well adapted to the greater part of the age table is as successful when applied to these lower ages. This follows from the fact that the error in the reports of children's ages is of a different nature from that in the reports of the ages of adults, and from the high rate of mortality among young children. There seems, therefore, to be no reason why knowledge respecting the ages of

[1] Cf. *Op. cit.*, or 'Censimento della Popolazione del Regno d'Italia al 31 Dicembre 1881'; Relazione Generale, p. 115.

[2] 'Census of England and Wales, 1891,' General Report, pp. 29, 105.

[3] Among the censuses which have followed this procedure are those of Germany, France, Hungary and Sweden.

children, gained from other sources, if of a more accurate character than is furnished by the census, should not be made the basis of the adjustment of the lower terms of the P_x series. For the United States as a whole such knowledge is lacking, although this process of adjustment could be successfully applied to the age returns of some of our state censuses.

II. Arbitrary Methods of Adjustment.

Under this head must be placed all attempts to smooth the age curve in which the justification for any particular change from the unadjusted figures is derived from the internal evidence yielded by an examination of the figures themselves. Such a method might be called analytical, were it not that the difficulties of a detailed analysis of the errors are so great that in applying the method one is compelled to rely very largely on more or less arbitrary assumptions. An interesting attempt of this kind was made by Mr. W. W. Drew, Superintendent of the Bombay Census of 1891. Mr. Drew's adjustment illustrates so well both the possibilities and the limitations of this kind of adjustment that his description of it will be quoted at length. After showing that the numbers especially favored in the age reports are, in order, the multiples of ten, the multiples of five, and the multiples of two, Mr. Drew says:[1]

> The conclusions that it seems to me to be quite fair to draw from this are that the ages not stated in round numbers are in the main correct; that where a person knows his age pretty thoroughly, but not the number of years quite accurately, he is more likely to record an even number than an odd one; that where he is more doubtful he will give the number ending in 5 or 0 that he thinks it to be nearest; and when he is quite uncertain he guesses at the nearest 10. For instance, a man of 43 who knew the exact day of his birth would record himself as such; but if he was not quite so accurate but knew his age pretty well he would be more likely to put himself down as 42 or 44 than 43. With still more uncertainty he would go to 45 or to 40; but he would not be likely to pass over the nearest number ending in 5, *viz.*, 45, for a more distant one like 35 or 55; nor 40

[1] 'Census of India,' 1891, Vol. VII, Bombay, Pt. 1, p. 55. For an account of an application of a similar method of adjustment to quinquennial age groups, see 'Census of India,' 1891. Assam, Vol. I. Report, p. 99.

for 30 or 50. A man of 46 would be almost equally likely to call himself 45 or 50; though a man of 49 would be much more likely to call himself 50 than 45. In other words, a number like 40 would attract entries from 36, 37, 38, 39, 41, 42, 43 and 44; 45 would attract from 42, 43, 46 and 47; 42 from 41 and 43; conversely, 43 would lose to 42, 44 and 45 and 40. I would therefore correct this table, by first, deducting from every age ending in an even number (including those ending in 0) the number of entries which would seem in excess, and dividing it between itself and the odd number on each side; secondly, in doing the same with each odd multiple of 5 (except 25), and dividing the excess between the number itself and the two nearest numbers on each side; and, thirdly, in doing the same with 25 and 12 and every even multiple of 5, and dividing the excess between the number itself and the eight numbers on each side of it. That is, I take the influence of the partiality for even numbers to affect only each even number itself and the odd number above and below it, that of each odd multiple of 5 to extend half-way between it and the next multiple of 5, and that for an even multiple of 5, which is the greatest, to extend half-way between itself and the next even multiple on either side, and to be greater on the numbers nearest it. From the even multiples of 10 two deductions have been made, one by which only the one number each side is benefited, as in the case of any even number, and one which extends half to itself and the two numbers on each side of it, on account of its being a multiple of 5 and half to itself and the four numbers on each side of it, on account of its being an even multiple of 5. In the case of a multiple of 10, to determine the excess attributable to each cause, I take the arithmetical mean between the entries on the even numbers immediately above and below it to show the entries that it would have borne, if it had been an ordinary even number. The excess above this is taken as the number of extra entries due to its being a multiple of 5. The way in which the excess on the ordinary even numbers has been divided, is by taking the arithmetical mean between the odd numbers on each side of it to be the number of entries recorded for it with the same accuracy as on the odd numbers, and subtracting these from the total on the even number, and then dividing the remainder between all three numbers, in the proportion of the amounts already on the odd numbers, and double that on the even.

It cannot be doubted that a process such as described will considerably "smooth" the age curve. It is probable that the results obtained will be appreciably nearer the truth than are the unadjusted figures. But that the process is arbitrary, rather than anlytical, is very evident. Mr. Drew has assumed that the tendency to over-statement of age is as strong as the tendency to under-statement. If, as is very probable, this assumption is erroneous, the entire process is vitiated. The number of terms depleted by the concentra-

tion on multiples of two, five or ten is the subject of another arbitrary assumption.

It does not seem that methods of this kind are desirable, even in such unique conditions as are furnished by the very erroneous census returns of India. If our knowledge of the form of error is sufficient to enable us to make a satisfactory adjustment by a process similar to that described above, such knowledge can be used to better advantage by being made the basis of a more objective method of adjustment.

III. Adjustment by the Use of Averages.

Under this head may be classed all methods of adjustment in which for any value of P_x there is substituted the average or mean value of the terms in a group of which P_x is the central or median term. In its simplest form such an adjustment may be expressed by the following formula, in which P_x is the number to be adjusted, and n is the number of terms in the group:

$$P'_x = \tfrac{1}{n}\left[P_{x-\frac{1}{2}(n-1)} \ldots\ldots + P_{x-2} + P_{x-1} + P_x + P_{x+1} + P_{x+2} \ldots\ldots + P_{x+\frac{1}{2}(n-1)}\right] \qquad \text{(A)}$$

Here the adjusted value is the simple arithmetic mean of the values of P_x included in the group. The simplest way of performing such an adjustment is to take the arithmetic mean of the values of P_x and the two neighboring terms; that is, to make

$$P'_x = \tfrac{1}{3}(P_{x-1} + P_x + P_{x+1}) \qquad \text{(B)}$$

In this case equal weight is given to each of the terms used in the adjustment. In practice such formulas are usually so constructed that different "weights" are given to the different terms, on the theory that terms which are near to the term to be adjusted should have more influence in determining its value than terms which are farther removed. It is interesting to note in this connection that the effect of adjusting a series by formula (B) and then readjusting the results by the same formula is equivalent to the effect of one application of the following formula:—

$$P'_x=\tfrac{1}{9}(P_{x-2}+2P_{x-1}+3P_x+2P_{x+1}+P_{x+2})$$

which may be written:

$$P'_x=\tfrac{3}{9}P_x+\tfrac{2}{9}(P_{x-1}+P_{x+1})+\tfrac{1}{9}(P_{x-2}+P_{x+2}) \qquad (C)$$

In formula (C) the weight decreases with the distance of the term from P_x. It is evident that these weights may be chosen arbitrarily and that there are an indefinite number of possible weights. Various rules have been proposed for choosing the weights that will give the best results. The problem is to find the best values of the coefficients c_0, c_1, c_2, etc., for the adjustment of a group of n terms by the formula:

$$P'_x=c_0P_x+c_1(P_{x-1}+P_{x+1})+ \quad . \quad . \quad . \\ +c_{\frac{1}{2}(n-1)}\left(P_{x-\frac{1}{2}(n-1)}+P_{x+\frac{1}{2}(n-1)}\right) \qquad (D)$$

The earliest adjustments by formulas of this nature were of the simplest kind. Mr. J. Finlaison, an English actuary, in a report on the Law of Mortality of the Government Life Annuitants (March, 1829), described a method used by him in the graduation of mortality experience. He used a formula like (B) above, but consisting of five terms in place of three. Applying such a formula he obtained for his adjusted value:

$$P'_x=\tfrac{1}{25}[5P_x+4(P_{x-1}+P_{x+1})+\ldots.+(P_{x-4}+P_{x-4})]$$

This formula is identical with that recommended by Th. Wittstein, a German actuary, nearly 40 years later.[1] Filipowski's method is very similar.[2] Assuming that $P_{x+\frac{1}{2}}$ should equal $\frac{1}{2}$ (P_x+P_{x+1}) he deduced the following formula:

$$P'_x=\tfrac{1}{4}(P_{x-1}+2P_x+P_{x+1})$$

In the formulas which have been given thus far there is one peculiarity to be noted. The principle on which they all are based is that in a series accurately representing the facts each term would be the arithmetic mean of the terms in any group of which it is the median term. Now this re-

[1] Wittstein, '*Mathematische Statistik*' (Hanover, 1867), p. 30; Journal of the Institute of Actuaries, XVII, pp. 418-420.

[2] Filipowski described his method in the 'Insurance Record' of December 9, 1870.

lation holds true only in a series of the first order, that is, a series whose law is arithmetical progression. Such series are represented graphically by straight lines. Thus the effect of applying one of these formulas to either an age or a mortality curve would be to straighten it as well as to smooth it. If the unadjusted curve is so irregular as to require several applications of the adjustment formula, the error introduced by this straightening process will be considerable. The values of P_x thus obtained will be too large when the curve is convex toward the axis of abscissas and too small when it is concave.

However, we are not limited in our choice of adjustment formulas to those which presuppose a series of the first order. We can assume that the facts may be represented by a curve of any given order,[1] and construct our adjustment formulas accordingly.

The general principles underlying the choice of these "weights" or coefficients seems to have been first considered by Schiaparelli, of Milan, in a study of the adjustment of meteorological observations.[2] It is not necessary to enter into a detailed discussion of the mathematical principles involved. The problem which Schiaparelli attacked may be stated as follows: Given a group of *n* terms, to find a combination of these terms that will represent the median term exactly, if the values of the terms are exactly known, and which will give the closest approximation to the value of the median term when the observed values of the terms are subjected to error.

Schiaparelli gives tables of the probable errors of the adjusted terms for series of different orders, but of the general parabolic form:

$$P_x = A + B_x + C_x^2 + \dots \text{etc.}$$

1 It is not necessary that a curve of the order chosen should represent the facts for the entire series It is sufficient if it fits the conditions of every group of n terms.

2 'Sul modo di ricavare la vera expressione delle leggi della natura delle curve empiriche,' [Milan, 1867].

The coefficients of the various terms for curves of different orders are easily obtained by substituting in the general formula (D) values derived from the principle that in a curve of the *n*th order, the *n*th differences are constant. The peculiar interest of Schiaparelli's work to the student of age statistics lies in the fact that it was the basis of several adjustments of the age reports of the Italian census of 1881. An account of these adjustments may be found in the studies already mentioned in our account of adjustment by substitution.[1] The simplest form of adjustment was used, the formulas being

$$P'_x = \tfrac{1}{3}(P_{x-1} + P_x + P_{x+1})$$

$$\text{and } P'_x = \tfrac{1}{5}(P_{x-2} + P_{x-1} + P_x + P_{x+1} + P_{x+2})$$

These formulas, as we have stated above, are accurate only when applied to a series of the first order. Now, whatever may be the form of the age curve, it is certainly not a straight line. To obtain the final adjustment in the case mentioned, the three term formula was applied twice to the crude returns in quinquennial groups. Values for single years were then interpolated, and the results were in turn smoothed by applying the three-term formula twice and the five-term formula three times. The resulting series is certainly "smooth," but how closely it represents the age constitution of the population at the time of the census is an open question. A graphic representation of the adjusted curve shows that it approaches the form of a straight line. The adjustment does not conform to one of the tests of a good adjustment—the general agreement of corresponding groups of terms of the adjusted and unadjusted series.

The justification given in the studies under consideration for the use of these simple forms of the adjustment formula is that the probable errors, as determined by Schi-

[1] *Annali di Statistica*, "*Sulla Composizione della Popolazione per Eta*," Rome, 1885. Cf. especially pp. 61 and ff. A short account of the method of adjustment used, together with some of the results, will be found in "*Censimento della Popolazione del Regno d'Italia al 31 Dicimbre, 1881*," *Relazione Generale*, pp. xxxix-xliv and 113-115.

aparelli, were less for two successive applications of a formula of the first order than for one application of a formula of the second order.[1] It is hard to see what this fact has to do with the matter. Schiaparelli deduced his coefficients of error on the assumption of an agreement between the order of the series and the adjustment formula. That is, he assumed that the adjustment formula was fitted to the form of the curve to be adjusted. If we apply a formula fitted for the adjustment of a series of the first order to a series of the second or third order, we introduce an error much greater than the difference between the "probable errors" of two adjustment formulas of the same form but of different orders. In fact, it is very certain that successive groups of terms of an accurate age series cannot be represented by curves of either the first or second degree. Such curves cannot contain points of inflexion, which, as we have seen, are usually present in age curves. Against the theory of adjustment upheld in the Italian age studies, it may further be stated that the errors in age reports cannot be treated as errors of observation. This point will be treated at greater length in another part of this study.

Some important contributions to the theory of the adjustment of irregular series were made by E. L. De Forest, whose first article on the subject appeared in the 'Smithsonian Report' for 1871.[2] De Forest devoted considerable study to different methods of obtaining the weights of coefficients used in adjustment formulas, and reached some very interesting results. He first used formulas in which the weights increased in arithmetical progression from the extreme terms to the middle one. These were discarded for formulas in which the coefficients formed a curve of the sixth order tangent to the axis of X at the zero weights,—that is, at the first terms not included in the formula. An-

1 *Loc. cit.*, p. 75.

2 For other articles by De Forest on this subject see *The Analyst*, Vols IV and V, also a pamphlet on 'Interpolation and Adjustment of Series,' New Haven, 1876.

other method which gave good results consisted in finding those values of the weights which would make the probable value of the fourth differences a minimum. This method assumes that the mortality curve (which De Forest had especially in mind) may be closely represented by successive curves of the third order. As such curves admit a point of inflexion, the assumption does not seem to be unfounded. De Forest suggested several other ingenious methods of obtaining adjustment coefficients. In his subsequent articles in 'The Analyst' these methods were perfected. It must be said that De Forest's studies are the most thorough that have been made with reference to this kind of adjustment. However, his methods are not especially applicable to census data and probably have never been used for adjusting age tables.

The next method to be considered under this head is one devised by Mr. W. S. B. Woolhouse, and first used by him in graduating the H_m table for the institute of Actuaries. His description of it is as follows:[1]

"If we begin at the first age in the table and extract the numbers living at quinquennial intervals, P_{10}, P_{15}, P_{20}, P_{25} . . . we can, by the formula for interpolation, determine all the intermediate values at the other ages, and so obtain a complete series of values that shall be continuous. Geometrically speaking, we shall then pass a continuous curve-line through the indicated quinquennial points. Against the adoption of such curve lines as the basis of the final table, there is manifestly this tangible objection, that the numbers at the ages of 10, 15, 20, 25, ... are made use of exclusively, and that the original numbers between those ages are wholly ignored as data. This rather material objection, which is inherent in other methods of adjustment, is entirely removed by varying the epoch of the adopted quinquennial data, that is, by taking the five distinct series hereunder stated, viz:

$$\begin{array}{llll} P_{10} & P_{15} & P_{20} & P_{25}\ldots\ldots \\ P_{11} & P_{16} & P_{21} & P_{26}\ldots\ldots \\ P_{12} & P_{17} & P_{22} & P_{27}\ldots\ldots \\ P_{13} & P_{18} & P_{23} & P_{28}\ldots\ldots \\ P_{14} & P_{19} & P_{24} & P_{29}\ldots\ldots \end{array}$$

Then by separately interpolating the intermediate values for each of these series, and by finally taking the arithmetical average or mean value, of the five completed sets of results. The logical

[1] *Journal of the Institute of Actuaries*, XV, pp. 390, 391.

premise that virtually guides us to this last deduction is the recognized principle, that the probabilities of positive and negative errors are equal. Reverting again to a graphical illustration, all the points of the original data are thus occupied by five distinct curves, assimilating to the experience and to one another, and forming in combination a kind of net work; and at every age the resulting ordinate of the adjusted curve is the arithmetical mean of the five corresponding ordinates, and the five curves are as it were mutually drawn in towards a central course."

It may seem that this method is in one sense a method of interpolation, rather than a method of averaging. But its essential principle is that the value of the adjusted term is the mean of the value of five other terms. Moreover, Mr. Woolhouse has moulded the rather cumbersome process described above into the application of a formula similar to those described by Schiaparelli and De Forest. Thus:

$$P'_x = .20P_x + .192(P_{x-1} + P_{x+1}) + .168(P_{x-2} + P_{x+2}) + .056(P_{x-3} + P_{x+3}) + .024(P_{x-4} + P_{x+4}) - .016(P_{x-6} + P_{x+6}) + .024(P_{x-7} + P_{x+7})$$

This formula seems to be the best of its kind, and for many years was standard among actuaries. Mr. Woolhouse's use of quinquennial intervals gives his method certain advantages in its application to census figures.[1] If the percentage of reported ages concentrated on multiples of 5 were the same in each quinquennial group, and if the other errors of statement were distributed proportionately on the first, second, third, fourth, and fifth ages of each quinquennial group, the Woolhouse method would give a very close approximation to the facts. Of course, the errors in an age table are not distributed so evenly, but there is enough correspondence between the actual conditions and the hypothetical conditions just considered to afford some justification for the use of the Woolhouse method.

There is, however, one important objection to the use of what we have called "methods of averages" in the ad-

[1] Mr. Woolhouse did not have census problems in mind when he evolved his formula. He states that his reason for adopting an interval of five terms had reference chiefly to ease of computation. [*Jour. Inst. Act.*, XX1X, 237.]

justment of census age returns. All of the adjustment formulas considered under this head were constructed with a view to the elimination of accidental errors. By accidental errors we mean natural errors arising from paucity of observations; errors whose distribution may be assumed to be effected by chance. With accidental errors the principle holds good that errors of a given amount are as liable to occur in one term as in another. That a given term has a positive error affords no basis for presuming that the error of the next term will be negative.

The errors in the age table are not of this nature. They are systematic errors and take certain definite forms. There are undoubtedly some accidental errors in the age tables, but they are few in number, and tend to neutralize each other. Schiaparelli's formulas were intended for the reduction of meteorological observations, while those of Woolhouse and De Forest were intended for the graduation of mortality tables based on the experience of life insurance companies. The errors in these tables are mainly such as would be eliminated if the number of observations were indefinitely large, and the application of adjustment formulas based on the general theory of errors is entirely justifiable.[1]

Physical observations, such as those made in astronomical research are often subject to both kinds of errors. Before an astronomer can apply the "method of least squares" with a view to obtaining the most probable value of a number of observations he must first eliminate the known errors. Until the known errors are eliminated, the law of error is not applicable, as it relates solely to residual errors.[2]

[1] It is stated in the Census of Assam (India), for 1891 [Vol. I, p. 97], that the errors in census age reports "tend to eliminate each other" and that "the approach to absolute correctness" varies "in direct proportion with the number of persons included in the returns." This statement is without foundation.

[2] Cf. Sorley, 'Observations on the Graduation of Mortality Tables,' *Jour. Inst. Act.*, XXII, 311.

It should also be remembered that the errors in census age tables are usually very large, while the adjustment formulas are adapted only to comparatively small errors. In order to smooth the age curve, it would usually be necessary to make several applications of any of these formulas. The necessary difference between the actual form of the age curve and the form assumed in the construction of any formula which might be used, would in this way introduce a considerable element of error. This point need not be developed farther, as it has already been brought out in the discussion of adjustment by simple arithmetical averages.

As has been suggested, the Woolhouse formula happens to be so constructed that its use in the adjustment of census age returns can be in a measure justified. But as Mr. Woolhouse himself said in the passage quoted above, it is based "on the recognized principle that the probabilities of positive and negative errors are equal." Since this hypothesis does not apply to the periodic errors of the census age tables, the use of the Woolhouse formula cannot be recommended. What we want is a method of adjustment based, not on the theory of errors, but on the peculiar conditions of the problem under consideration.

It has seemed necessary to discuss this matter at some length, for in at least three instances, methods of adjustment based on the theory of errors have been applied either to the census age returns or to mortality tables based on such returns.[1] Of course what has been said against the use of these adjustment formulas for the smoothing of the age curve applies with equal force against their use in smoothing mortality tables based on census statistics.

Methods of Algebraic Interpolation.

From a mathematical standpoint, interpolation is not a method of adjustment, but a method of obtaining values

[1] Cf. besides the Italian studies already mentioned, 'Census of New South Wales,' 1891, Statisticians' Report, p. 150; 'Eighth Census of the United States,' 1860, Mortality and Miscellaneous Statistics, p. 518.

of the intermediate terms of a series from known values taken at fixed intervals. The ordinary method of interpolation is by the use of finite differences, although in theory any method by which the values of the constants in an assumed equation can be obtained may be used as a method of interpolation.[1]

In the special problem presented by the census age returns, the particular method of interpolation is less important than the method of securing the "known values at fixed intervals," since none of the terms of the P_x series can be supposed to be accurate. In the adjustment of the age returns of the English census, this difficulty is surmounted by graduating the Q_x instead of the P_x series. (Q_x represents the number living at and above any age x.) It has been shown that the value of a group is more likely to be in accordance with the facts than is the value of a single term. The difference in the accuracy of groups and of single terms is more marked when the form of group used is chosen with reference to the nature of the error in the reported ages.

Thus it is assumed in the case of the English census that the most accurate groups are the decennial groups P_5—P_{15}, P_{15}—P_{25}, P_{25}—P_{35}, etc., in which the year of greatest concentration is the median year. Now the values of these groups are the differences between the values of Q_x taken at corresponding decennial intervals, and the error in the Q_x series is no larger[2] than the error of the series formed by the decennial groups. Here then, we have "known values at fixed intervals" which may be conveniently used for interpolation. It should be noted that

1 De Forest, in his article in the 'Smithsonian Report' of 1871, already mentioned, develops a method of interpolation based on the principle that "in a continuous series whose law is given or assumed, the sum of a limited number of terms can be regarded as a definite integral, which is the aggregate of a succession of similar integrals corresponding to the terms considered" [*Loc. cit.*, p. 277.]

When the known terms of a series are subject to errors of observation, the "method of least squares" is most advantageous.

2 The relative error is less.

the terms of the P_x series are the first differences of the Q_x series. In interpolation by finite differences the first differences of the adjusted series cannot be expected to progress as smoothly as the series itself. This difficulty is partly overcome in the practice of the English census by interpolating the values of *log* Q_x.[1]

This method of adjustment seems better adapted to census figures than any of the other methods thus far examined. But the form of age groups used in the English practice seems open to criticism. In the first place a group of ten terms is too large. It covers up too many of the real irregularities of the age series. Moreover, the placing of the year of greatest concentration in the center of the group probably retains a greater error in the series than would be present if the first term of each group were a year of concentration. This follows from the fact that analysis of age returns shows that many more persons under-state their ages than over-state them.[2]

In the appendix to this article will be found a table showing the results of applying this method of interpolation to the returns of the United States census of 1890. The process used was as described above, except that the fixed values used were log Q_{10}, log Q_{15}, log Q_{20}, etc. It will be seen that the P_x curve is fairly smooth, with the exception of a few points at the junctures of the quinquennial groups. These are due to the fact that the irregularities in the progression of the groups are reflected in the first differences of the Q_x series.[3]

1 By using the logarithm a constant *percentage* of error is changed to a constant *amount* of error.

2 Cf. my article on "The Comparative Accuracy of Different Forms of Quinquennial Age Groups," *Quarterly Publications of the American Statistical Association*, March, 1900.

3 The adjustment was made by the use of four orders of central differences. For a good description of various methods of applying the principle of finite differences to interpolation, see the chapters on 'Finite Differences' and 'Interpolation' in the 'Text Book of the Institute of Actuaries,' Vol. II. See also Rice, 'Theory and Practice of Interpolation,' Boole, 'Finite Differences,' and Bowley, 'Elements of Statistics,' p. 242 and ff.

The only other method of interpolation that needs consideration in this connection is interpolation by an exponential formula; more specifically, by the Gompertz-Makeham law of mortality. This formula may be written:[1]

$$P_x = k a^x b^{c^x}$$

in which k, a, b, and c, are constants determined from known terms of the series. This formula has been much used in actuarial practice, and has been found to express the decrease in the numbers living from the age of 10 upwards with a close degree of approximation.

It is evident that no formula of this kind can be expected to represent the age constitution of an actual population. It takes account of only one of the factors which determines the age constitution of the population. This formula is mentioned here because it was used by Mr. G. F. Hardy in the graduation of the age tables of the census of India of 1881.[2] Mr. Hardy retained the results given by the formula only for the ages above 60. For the lower ages his results were used "simply as a *base line* by which to adjust, by the graphic method, the actual numbers recorded in the Census Returns, the adjusted curve being made to run into this base line about age 60."[3] As Mr. Hardy's adjustment was made for the purpose of constructing life tables, which are intended to represent only the normal mortality, his procedure seems entirely justifiable. But for the purpose of obtaining a true statement of the age constitution of an enumerated population arbitrary formulas of this nature are useless; unless, indeed, the number of constants in a formula is made equal to the number of known values which it is desired to retain in the adjusted curve. In such cases, however, an arbitrary formula loses any peculiar

1 This formula is based on the supposition that there are two components in the law of human mortality: one expressing a chance distribution of deaths; the other expressing the decrease (with advancing age) of the ability to withstand the forces tending to close life.

2 Report on the Census of British India taken on the 17th February, 1881. Vol. I, pp. 160-162.

3 *Loc. cit.*, p. 161.

claim to represent a "law of mortality" and becomes a rather cumbersome method of interpolation.

V. Graphic Methods.

If an irregular series be represented graphically, and if a smooth curve be drawn which follows the general form of the irregular curve as closely as possible, we have an example of adjustment by the graphic method. If the terms of an irregular series are in successive groups, these groups may be represented by successive paralellograms, the width of each paralellogram being fixed by the number of terms in the corresponding group, and its height by the average value of the terms. Then if a smooth curve is drawn through the upper ends of the paralellograms, in such a manner as to add to each group as much area as it cuts off, we have an example of interpolation by the graphic method.[1]

This method has been used in several censuses. Its advantages are manifest. It can be made to give as smooth a curve as is desired, and its application requires no mathematical knowledge. On the other hand it has the manifest disadvantage that no two persons in applying it would obtain exactly the same results. If any one using the method had preconceived ideas as to the form of the age curve, they would very probably be reflected in his results. Moreover, the small scale on which the curve must in practice be drawn, prevents a close reading of the values of the interpolated terms. Probably a very good curve could be obtained by readjusting the results of a graphic adjustment by means of the Woodhouse formula. The same considerations as to the form of age groups to be used apply to the graphic method as to the method of differences.

[1] The best description of the graphic method, although from the standpoint of a partisan, is by T. B. Sprague, 'The Graphic Method of Adjusting Mortality Tables,' *Jour. Inst. Act.*, XXVI, pp. 77-112. See also W. S. Jevons, 'Principles of Science,' Sec. on 'The Graphical Method.' It may be of interest to note that the graphic method was used by Milne in the graduation of his famous Carlisle life tables. See the articles by Wm. Sutton and Geo. King on the 'Method used by Milne in the Construction of the Carlisle Table of Mortality' in Vol. XXIV of the *Jour. Inst. Act.*

General Considerations.

It seems that of the methods discussed, interpolation by differences and the graphic method possess certain advantages, when the curve to be smoothed is a census age table. Of the two, the graphic method is the more flexible, and will give possibly a smoother curve. It would appear that these advantages are more than balanced by the fact that the method of interpolation by differences is more definite and objective. It will give identical results if applied independently by two or more persons.[1] The principles underlying interpolation by differences can be readily comprehended by anyone whose knowledge of algebra includes the binomial theorem.

Both the graphic method and interpolation by differences give results which can be arranged in age groups whose totals will correspond to the totals of the corresponding groups of enumerated ages.

The graphic method fails completely at the more advanced ages. The values of the groups of terms above P 85 are so small that they do not afford a sufficient basis for a graphic adjustment. The method of differences is more successful, but does not give as good results as at the lower ages. The percentage of error in the reported ages above 80 or 85 is very high, and the error is of a different nature from that found in the reports of lower ages. Persons above 80 or 85 systematically over-state their ages. If the adjustment of these higher ages is necessary, it could be done best perhaps by using the decennial groups 85—94. 95+, which are probably more accurate (for these ages) than the quinquennial groups recommended for the lower ages. For ages below 20 there is no perceptible concentration on multiples of five. Hence it is probably sufficient to begin the adjusted series with P_{20}. It may be added that no method of adjustment is satisfactory when applied to ages

[1] Assuming, of course, that the different computers use the same orders of differences and the same formula.

below five. The error here also takes the form of overstatement. This, coupled with the high mortality, makes it difficult to improve upon the reported figures, except by substituting figures drawn from more accurate sources.

ALLYN A. YOUNG.

APPENDIX.

AGES OF THE POPULATION OF THE UNITED STATES; CENSUS OF 1890.

Age.	Number Reported as of Each Age.	Adjusted Number.	Age.	Number Reported as of Each Age.	Adjusted Number.
20	1,282,322	1,295,403	50	776,333	516,735
21	1,246,876	1,272,509	51	336,202	492,880
22	1,275,042	1,244,038	52	440,347	466,542
23	1,225,888	1,210,876	53	387,734	438,966
24	1,166,548	1,173,850	54	385,646	411,139
25	1,173,342	1,112,412	55	437,032	363,208
26	1,041,110	1,076,307	56	375,254	345,148
27	979,887	1,042,899	57	305,830	331,280
28	1,142,216	1,012,159	58	313,340	320,579
29	891,222	984,000	59	240,880	312,121
30	1,359,566	969,057	60	502,788	321,397
31	729,771	942,977	61	206,016	308,576
32	908,090	916,334	62	261,577	293,378
33	816,613	889,084	63	256,730	276,416
34	764,590	861,178	64	230,923	258,267
35	1,013,609	832,495	65	310,320	230,578
36	770,655	802,57	66	195,990	215,192
37	673,381	772,768	67	183,170	201,066
38	789,875	743,441	68	181,546	187,885
39	618,641	714,882	69	139,084	175,389
40	1,037,336	682,948	70	245,007	168,122
41	486,853	658,092	71	110,117	154,187
42	630,022	635,247	72	132,706	140,135
43	533,183	614,261	73	113,126	126,314
44	498,124	594,970	74	100,795	112,993
45	779,816	575,016	75	122,098	97,373
46	524,565	560,405	76	85,204	87,175
47	468,635	546,386	77	65,702	77,892
48	533,040	532,279	78	71,032	69,313
49	425,584	517,554	79	49,026	61,309